**Project Requirement and Specification**

**on**

OUTLIER DETECTION

**(CSE 5th Semester Mini Project )**

2021-2022



**Guided by: Submitted by:**

Mrs. Parul Madan Yash

(Resource Person) (2015200)

DS & AI

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# GRAPHIC ERA DEEMED TO BE UNIVERSITY, DEHRADUN

# ACKNOWLEDGEMENT:

I would like to express my sincere gratitude to Mrs. Parul Madan mam of the department of Computer Science, whose role as project guide was invaluable for the project. I am extremely thankful for the keen interest she took in advising me. I convey my gratitude to all the teachers for providing us the technical skill that will always remain as my asset. Last but not the least, I wish to thank my parents for financing my studies in this college as well as for constantly encouraging me to learn new things.

**MOTIVATION:**

Machine Learning is a part of Data Science and I am being interested in everything having a relation with the Machine Learning, the independent project was a great occasion to give me the time to learn and confirm my interest for this field. The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities. We can use Machine Learning in almost everywhere. That’s why I decided to conduct my project around Machine Learning.

**CONTEXT:**

This project has been part of my 5th semester of Bachelors of Technology CSE course offered by the esteemed Graphic Era University. As part of the Mini project, I chose to make a project based on outlier detection with the help of few machine learning methods.

**ABOUT PROJECT:**

My project is all about detecting the Outliers from the given dataset.

What are Outliers?

In machine learning Outliers are nothing but the extreme values that do not follow the pattern in the data or the value that diverge from all the other values.

These can arise due to human error, instrument error or simply through natural deviations in populations.

Problems caused by Outliers?

It increases the error variance. They can cause bias and/or influence estimates. They can also impact the basic assumption of regression as well as other statistical models. These affect the mean and standard deviation of a dataset.

We can either drop the outliers or we can keep them as per the given circumstances.

* Examine an outlier if it affects your result.
* The important point is that if you drop the outliers then trim the dataset and replace the outliers with the nearest “good” data.

Methods for detecting Outliers.

1. Z-score method.
2. IQR (Inter Quartile Range) method etc.

**Z-score**

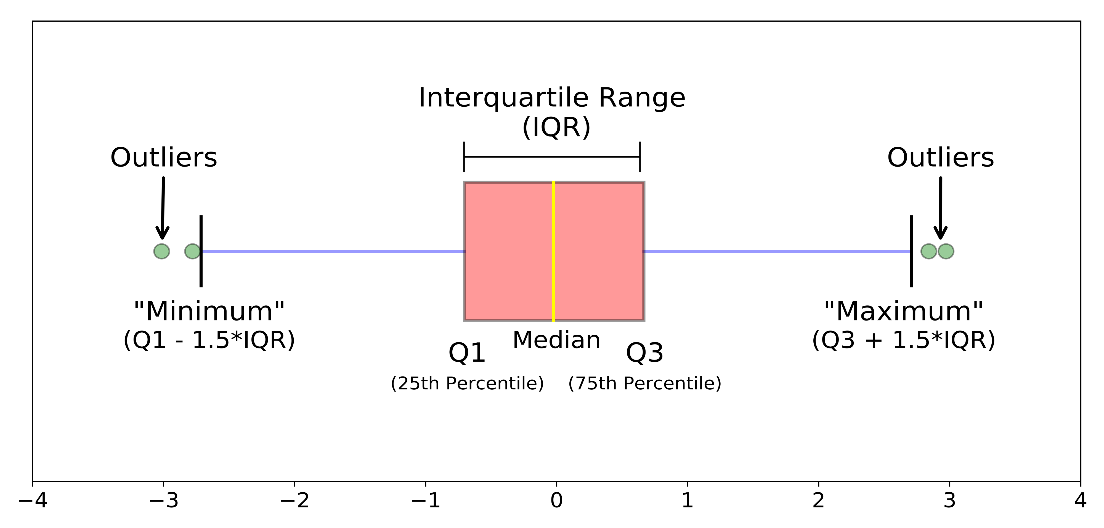
Using Z score method, we can find out how many standard deviations value away from the mean.

Z- score formula = (X-Mean) / Standard deviation

If the z score of a data point is more than 3 (because it cover 99.7% of area), it indicates that the data value is quite different from the other values. It is taken as an outlier.

**IQR (Inter Quartile Range)**

IQR tells us the variation in the data set. Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR treated as outliers. I have used this method in my project.



\* Q1 represents the 1st quartile/25th percentile of the data.

\* Q2 represents the 2nd quartile/median/50th percentile of the data.

\* Q3 represents the 3rd quartile/75th percentile of the data.

\* (Q1–1.5\*IQR) represent the smallest value in the data set and (Q3+1.5\*IQR) represent the largest value in the data set.

In my project I have taken customers dataset.

We should make sure that the data is already sorted before finding the inter-quartile range so I have sorted the income and spending values.

After calculating the 25th and 75th percentile of the given data, I have calculated the upper and lower bound here upper and lower bound refers to the maximum and minimum value of the IQR.

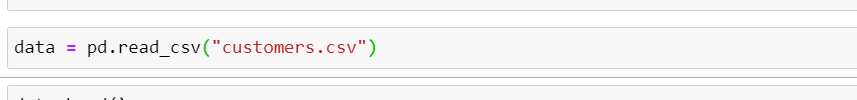
Any value greater than the upper bound and less than the lower bound will be considered as the Outlier.

This method is used for skewed distribution.

**LIBRARIES USED:**

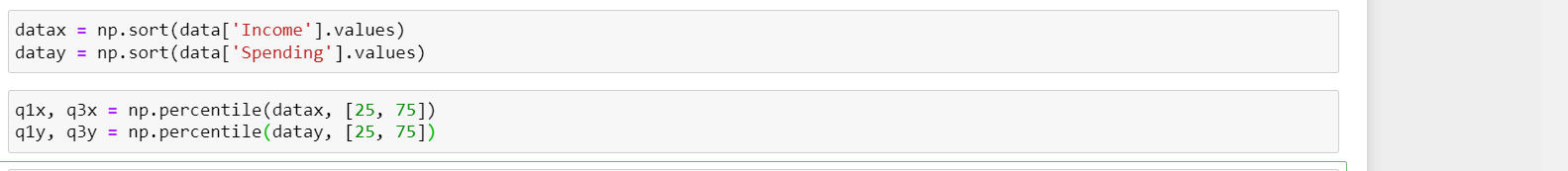
**Pandas :**- Pandas is a software library written for the Python programming language for data manipulation and analysis.

I have imported pandas library as pd which helped me in reading my csv file (dataset) into the code.



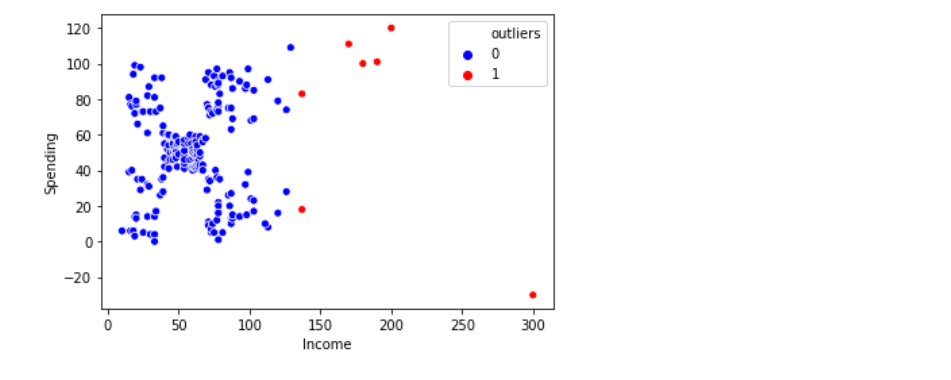
**Numpy :-** It is a library used for working and performing mathematical

operations on arrays. I have imported this library as np which helped me in sorting my data and also calculating the 25th and 75th percentile of the data.



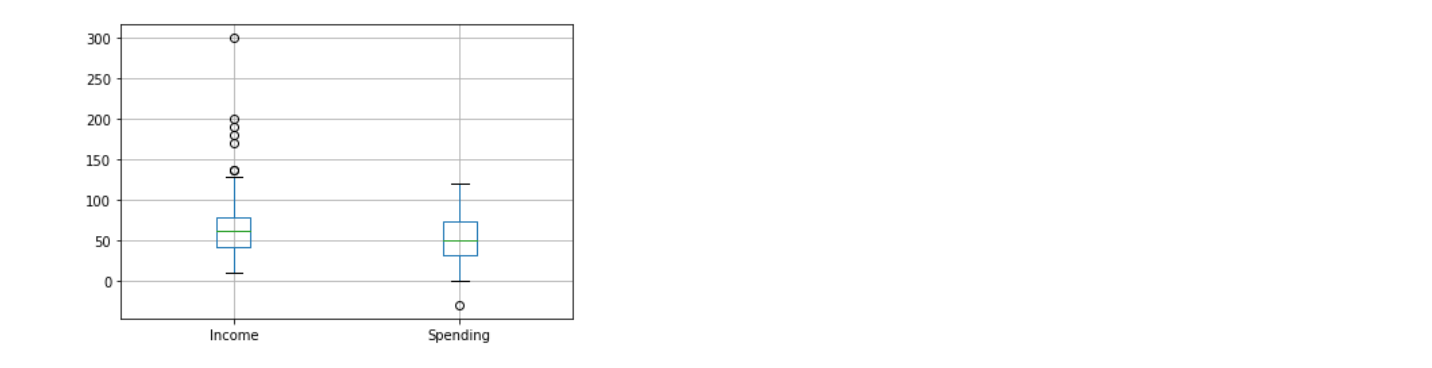
**Seaborn :-** It is an open-source Python library built on top of matplotlib used for data visualization and exploratory data analysis.

I have imported seaborn library as sns with the help of which my final scatterplot with the outliers is represented.



Here the red dots are the outliers which crossed the maximum and minimum values of the IQR range.

Also, I am able to show the boxplot representation with the help of seaborn library



The main objective of my project is to detect the outliers in the given dataset which I have done using the IQR method which differentiate the data points.

**CONCLUSION:**

Outlier detection is extensively used in a wide variety of applications such as intrusion detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems and in various kind of images.

With this project I have learnt about few Machine Learning methods which includes Z score method, IQR method and DBSCAN algorithm also.

**REFERENCES:**

Websites:

<https://towardsdatascience.com/>

<https://www.kaggle.com/>

Books:

Machine learning by Tom M. Mitchell

## [Machine Learning in Action by Peter Harrington](https://www.amazon.com/gp/product/1617290181/)